Walks on SPR Neighborhoods

Alan Joseph J. Caceres*† Juan Castillo*† Jinnie Lee* Katherine St. John*

October 7, 2011

Abstract

A nearest-neighbor-interchange (NNI) walk is a sequence of unrooted phylogenetic trees, $T_0, T_1, T_2, \ldots$ where each consecutive pair of trees differ by a single NNI move. We give tight bounds on the length of the shortest NNI-walks that visit all trees in an subtree-prune-and-regraft (SPR) neighborhood of a given tree. For any unrooted, binary tree, T, on n leaves, the shortest walk takes $\theta(n^2)$ additional steps than the number of trees in the SPR neighborhood. This answers Bryant's Second Combinatorial Conjecture from the Phylogenetics Challenges List, the Isaac Newton Institute, 2011, and the Penny Ante Problem List, 2009.

1 Introduction

Evolutionary histories, or phylogenies, are essential structures for modern biology [10]. Finding the optimal phylogeny is NP-hard, even when we restrict to tree-like evolution [8, 12]. As such, heuristic searches are used to search the vast set of all trees. There are many search techniques used (see [15] for a survey), but most rely on local search. That is, at each step in the search, the next tree is chosen from the "neighbors" of the current tree. A popular way to define neighbors is in terms of the subtree-prune-and-regraft (SPR) metric (defined in Section 2). For a given unrooted tree on n leaves, or taxa, the SPR-neighborhood is the number of trees that are differ by a single SPR move. The number of trees in the SPR neighborhood is $(2n - 6)(2n - 7)$. The second "Walks on Trees" conjecture of Bryant [5, 14] focuses on efficiently traversing this neighborhood via the nearest-neighbor-interchange (NNI) tranformations (defined in Section 2). Bryant asks:

> An NNI-walk is a sequence $T_1, T_2, \ldots, T_k$ of unrooted binary phylogenetic trees where each consecutive pair of trees differ by a single NNI.
>
> i. [Question] What is the shortest NNI walk that passes through all binary trees on n leaves?
>
> ii. [Question] Suppose we are given a tree T. What is the shortest NNI walk that passes through all the trees that lie at most one SPR (subtree prune and regraft) move from T?

Bryant's conjectures were posed as part of the New Zealand Phylogenetic Meetings' Penny Ante Problems [5] as well as the Challenges problems from the most recent Phylogenetics Meeting at the Isaac Newton Institute [14].

We answer the second question, proving that the shortest walk takes $\Theta(n^2)$ more steps than the theoretical minimum that visits every tree exactly once (that is, a Hamiltonian path). This builds on past work [6] that showed that a Hamiltonian path was not possible.

*Dept. of Math & Computer Science, Lehman College, City University of New York, Bronx, New York, 10468; Corresponding author: stjohn@lehman.cuny.edu; Support provided by NSF Grant #09-20920.

†Supported by an undergraduate research fellowship from the Louis Stokes Alliance for Minority Participation in Research, NSF #07-03449.

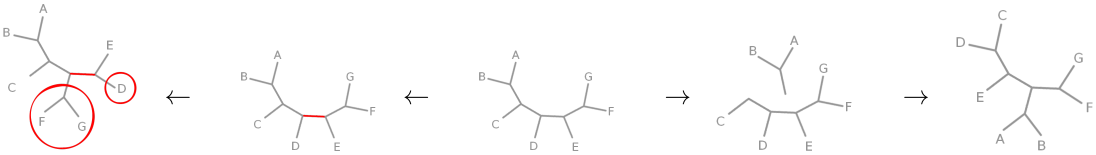

Figure 1: The trees on the left and center differ by a single NNI move. The tree on the right differs by a single SPR move from the center tree.

2 Background

This section includes definitions and results that we use from Allen & Steel [1]. For a more detailed background on mathematically phylogenetics, see Semple & Steel [13].

Definition 1. *An unrooted binary phylogenetic tree (or more briefly a binary tree) is a tree whose leaves (degree 1 vertices) are labelled bijectively by a (species) set S, and such that each non-leaf vertex is unlabelled and has degree three. We let $UB(n)$ denote the set of such trees for $S = \{1, \ldots, n\}$.*

Each internal edge, e of a tree $T \in UB(n)$ yields a natural bipartition, or **split** of the taxa; We write $A \mid B$ if there is an edge which partitions the leaf set into the two sets A and B. We use the standard notation of T_A to refer to the smallest subtree of T containing leaves only from A.

Figure 1 shows several binary trees. Each edge of a tree induces a **split** of the leaf set S. The *Nearest Neighbor Interchange* (NNI) is a distance metric introduced independently by DasGupta *et al.* [7] and Li *et al.* [11]. Roughly, an NNI operation swaps two subtrees that are separated by an internal edge in order to generate a new tree [1].

Definition 2. *Allen and Steel [1]: Any internal edge of an unrooted binary tree has four subtrees attached to it. A* **nearest neighbor interchange** *(NNI) occurs when one subtree on one side of an internal edge is swapped with a subtree on the other side of the edge, as illustrated in Figure 1.*

Definition 3. *The* **NNI distance**, *$d_{NNI}(T_1, T_2)$, between two trees T_1 and T_2 is defined as the minimum number of NNI operations required to change one tree into the other.*

The complexity of computing the NNI distance was open for over 25 years, and was proven to be NP-complete by Allen and Steel [1]. For a tree with n uniquely labeled leaves, there are $n-3$ internal branches. Thus, there are $2(n-3)$ NNI rearrangements for any tree.

One of the most popular moves used to search treespace is the Subtree-Prune-and-Regraft (SPR). Roughly, an SPR move prunes a selected subtree and then reattaches it on an edge selected from the remaining tree (see Figure 1).

Definition 4. *Allen and Steel [1]: A* **subtree prune and regraft** *(SPR) on a phylogenetic tree T is defined as cutting any edge and thereby pruning a subtree, t, and then regrafting the subtree by the same cut edge to a new vertex obtained by subdividing a pre-existing edge in $T - t$. We also apply a forced contraction to maintain the binary property of the resulting tree (see Figure 1).*

Definition 5. *The* **SPR distance**, *$d_{SPR}(T_1, T_2)$, between two trees is the minimal number of SPR moves needed to transform the first tree into the second tree.*

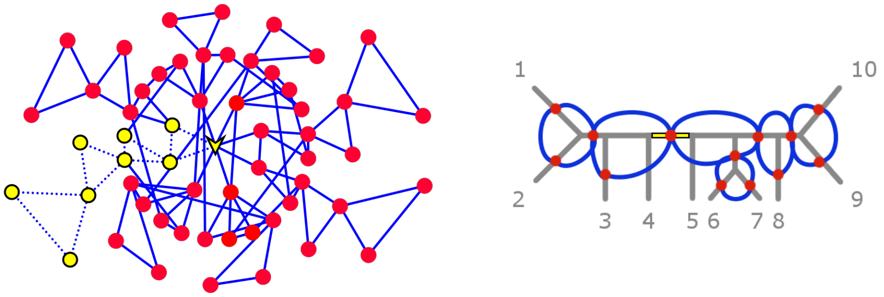

Figure 2: Left: The SPR-neighbor of a 7-leaf caterpillar tree. The highlighted nodes show the trees in the orbit that prunes a leaf from one of the sibling pairs. We note that any NNI-walk that visits every tree in this SPR neighborhood will visit some trees more than once. Right: The orbit of the edge, $e = 1, 2, 3, 4 \mid 5, 6, 7, 8, 9, 10$ from the tree $(1, (2, (3, (4, (5, ((6, 7), (8, (9, 10))))))))$ shown with respect to the tree. The tree is shown in the background with edge e highlighted. An SPR move is determined both by the edge pruned and the target edge of the regrafting. The trees in the orbit (red dots) are shown relative the regrafting edge in the initial tree. The blue lines indicate NNI-edges in the orbit. We note that the edges adjacent to e yield the initial tree when used as the target edge, so, do not produce a new tree.

The calculation of SPR distances has been proven NP-complete for both rooted and unrooted trees [4, 9]. Approximation algorithms for SPR on rooted trees exist [2, 3].

Definition 6. *Let T_0 be an unrooted, binary tree. Define $N_{SPR}(T_0)$ to be the* **SPR-neighborhood** *of T_0, namely,*

$$N_{SPR}(T_0) = \{T \mid d_{SPR}(T_0, T) = 1\}$$

When the tree in question is obvious, we will drop the argument and call the neighborhood N_{SPR}.

Definition 7. *Let T_0 be an unrooted, binary tree and and S be a set of trees that are 1 SPR move from T_0. Define $N_{NNI}(S, T_0)$ to be the* **NNI-neighbors** *of S, namely,*

$$\begin{aligned} N_{NNI}(S, T_0) = \quad & \{T \mid \exists T' \in S, d_{NNI}(T, T') = 1 \\ & \text{and } d_{SPR}(T_0, T') = 1\} \end{aligned}$$

We note that for every subset S of the SPR neighborhood of T_0, $S \subseteq N(S)$.

A "sibling pair" or "cherry" in a tree are two leaves that have the same parent. A "caterpillar tree" refers to the unrooted tree with exactly 2 sibling pairs.

3 Results

Theoretically, the shortest walk of the SPR neighborhood would be if each tree could be visited exactly once (that is, a "Hamiltonian path"). In [6], this was shown to be impossible for $n \geq 6$. This was done by showing that in the SPR neighborhood of caterpillar trees, there are at least 4 'isolated triangles' on the outer edge of the neighborhood (see Figure 2) that force at least two trees to be visited twice.

To bound the number of steps needed to visit the SPR neighborhood, we first introduce a new concept of an orbit of an edge. The orbit is all trees created by breaking that edge (in either direction) in an SPR move. More formally:

Definition 8. *Define for each edge e of the tree T_0, the* **orbit** *of e, O_e, to be all the trees that are one SPR move from T_0 where the edge broken by the SPR move is e.*

In Figure 2, the SPR-neighbohood, as well as orbits, of the 7-taxa caterpillar tree are shown.

To calculate the size of the SPR neighborhood of a tree, Allen and Steel (proof of Theorem 2.1, [1]) characterized the relationship between the trees in the neighborhood.

Theorem 9. *Allen and Steel [1]: Let T_0 be an unrooted phylogenetic tree on n leaves and let $N_{SPR} = N_{SPR}(T_0)$ be all trees that are a single SPR move from T_0.*

1. *The size of the SPR neighborhood is $|N_{SPR}| = 2(n-3)(2n-7)$.*
2. *The number of trees in N_{SPR} that can be obtained by more than one SPR move from T_0 are exactly those from the NNI transformations. Thus, there are $2n-6$ of them.*
3. *The number of trees in N_{SPR} that can be obtained by only one SPR move from T_0 is $4(n-3)(n-4)$.*

From this, we make the following simple observations about orbits:

Observation 10. *Let T_0 be an unrooted phylogenetic tree on n leaves:*

1. *Every tree $T \in N_{SPR}$ belongs to some orbit O_e for $e \in E(T_0)$.*
2. *Each orbit contains T_0.*
3. *Excluding T_0, there are exactly $2n-6$ trees that are included in two orbits.*
4. *The number of orbits is the number of edges in the tree, $2n-3$.*
5. *The size of each orbit is $2n-5$.*

The SPR neighborhood is the union of the orbits, but surprisingly, these orbits are mostly disjoint. Roughly, the overlap of orbits is very small and they have very few neighbors in common. Formally:

Lemma 11. *Let T_0 be an unrooted phylogenetic tree on n taxa. Let $T_1, T_2 \in N_{SPR}(T_0)$ and there exists $e \in E(T_0), T_1, T_2 \in O_e$. Let e_i be the target edge of the move that created T_i for $i = 1, 2$ (that is, T_1 is formed by grafting some pruned subtree of T_0 to e_1 and T_2 is the result of grafing a pruned subtree to e_2).*

Then, T_1 and T_2 differ by a single NNI move if and only if e_1 and e_2 have a common endpoint in T_0.

Proof. $\Longleftarrow$: Assume that e_1 and e_2 have a common endpoint in T_0. Let M be the subtree pruned by the SPR move that creates T_1. Without loss of generality, let the split induced by e_1 in T_0 be $ABC \mid DEM$ and the split induced by e_2 in T_0 be $AB \mid CDEM$, where A,B,C,D,E, and M are the leaves of subtrees of T_0. Let T_X refer to the subtree with leaves only from the set X.

If T_M is pruned to create T_1, then we have that T_1 contains the splits: $ABM \mid CDE$ and $AB \mid MCDE$. If T_M is also pruned to create T_2, then we have that T_2 contains the splits: $ABCM \mid DE$ and $ABC \mid MDE$. Thus, T_1 and T_2 differ by a single NNI move (swapping T_C and T_M), and the hypothesis holds.

So, assume that T_M is not pruned to create T_2, but instead that e is pruned in the other direction. Let $N = S - M$, where S is the set of leaves of T. Since e_1 and e_2 share an endpoint, at least one of them must be the edge pruned, e. If both are e, then $T_1 = T_2 = T$, and the hypothesis is trivially true. If only one, say e_2, is e, then e_1 must be a neighbor of e in T which implies $T_2 = T_1 = T$, and again the hypothesis is trivially true.

$\Longrightarrow$: By assumption T_1 and T_2 differ by a single NNI move. By definition of the NNI move, there exists an edge $e' \in E(T_1)$ that when removed, breaks T_1 into 4 distinct subtrees, T_A, T_B, T_C, T_D with leaf sets, A, B, C, D, and the split $AB \mid CD$ belongs to T_1 while $BC \mid AD$ belongs to T_2. Since both T_1 and T_2 are in the same orbit, the same edge e is pruned to create both. We note that since they differ by only the NNI move, that, by the argument above, the pruning of e must occur in the same direction for both to be result in nontrivial trees. Further, e must prune one of the subtrees: A, B, C, D, since only one move is allowed and T_1 and T_2 contain exactly the same trees. Without loss of generality, assume that A is pruned. We note the trees induced by the leaves of B, C, D are identical for these trees: $T_0|_{L(B \cup C \cup D)} = T_1|_{L(B \cup C \cup D)} = T_2|_{L(B \cup C \cup D)}$. It follows that e_1 and e_2 share a common endpoint, namely the intersection point of B, C, D. □

We can immediately give an upper bound on the NNI-walk of the SPR neighborhood as $O(n^2)$ steps. The underlying idea is to traverse each orbit separately, and then link these paths to form a traversal of the entire SPR neighborhood:

Lemma 12. *The SPR neighborhood has an NNI-walk of length $O(n^2)$.*

Proof. We will break the NNI-walk of the SPR neighborhood into a NNI-walk of the orbit of each edge in T_0. Since each orbit contains the initial tree T_0, we can glue together the walks of the orbit to make a walk of the entire space. We note that since each orbit contains at most $2n-5$, walking the $2n-3$ orbits in this fashion yields a walk with the number of steps is bounded by $2(2n-5)(2n-3) = O(n^2)$.

It suffices to show that there is a 2-walk of each orbit O_e for $e \in E(T_0)$. Each tree, $T \in O_e$, is created by pruning the edge e in T_0 and regrafting the pruned subtree to another edge in T_0 (see Figure 2). Every tree in the orbit corresponds to an edge in T_0 (namely, the target edge), and the trees in the orbit are connected exactly when their target edges share an endpoint in T_0 by Lemma 11. Thus, the orbit can be traversed by at most $2(2n-3)$ steps by starting at T_0 and following a depth-first-search of the tree (each tree in the orbit is visited at most once on the way "down" the search and once on the way "up" the search). □

To show the lower bound takes more work. It follows from this lemma that every orbit has very small overlap with the other orbits:

Lemma 13. *Let $T_1, T_2 \in N_{SPR}(T_0)$ such that T_1 and T_2 are a single NNI move apart. Then $d_{NNI}(T_0, T_1), d_{NNI}(T_0, T_2) \leq 2$.*

Proof. We note that if there exists an $e \in E(T_0)$ such that $T_1, T_2 \in O_e$, then by Lemma 11, the lemma holds.

So, let us assume that there exists $e_1, e_2 \in E(T_0)$, $T_1 \in O_{e_1}$, $T_1 \notin O_{e_2}$, $T_1 \in O_{e_1}$ and $T_1 \notin O_{e_2}$. Let M_i be the leaves of the subtree pruned with e_i from T_0 to create tree T_i, $i = 1, 2$. Since T_1 and T_2 are a single NNI move apart. By definition, there exists a split in T_1, $AB \mid CD$ that is rearranged in T_2: $BC \mid AD$.

We will argue, by cases, that both T_1 and T_2 are within 2 NNI moves of T_0. Without loss of generality, we will assume that $M_1 \cap T_A \neq \emptyset$

Case 1: $M_1 \subsetneq T_A$. Then, let $A' = A - M_1$. So, we have that T_1 contains the split $A'M_1B|CD$ and T_2 contains the split $BC \mid A'M_1D$. Since T_1 is only one SPR move from T_0, the structure of the 2 trees is identical without M_1, that is, $T_1|_{A' \cup B \cup C \cup D} = T_0|_{A' \cup B \cup C \cup D)}$, and T_0 includes and edge that splits A' and B from C and D. Since T_2 does not contain such an edge, the move that creates it must prune one of T_{M_1}, $T_{A'}$, T_B, T_C, or T_D. Pruning T_{M_1} is not possible since T_1 and T_2 are in different orbits. Pruning $T_{A'}$ is only possible if T_{M_1} and T_B have the same parent, in which case the $d_{NNI}(T_0, T_1), d_{NNI}(T_0, T_2) = 2$ and the lemma holds. Pruning T_B to create T_2 means that the subtree T_{M_1} is on the edge separating A' and B from C and D in T_0 and $d_{NNI}(T_0, T_1), d_{NNI}(T_0, T_2) = 2$. Lastly, pruning T_B, T_C, or T_D is only possible if $T_0 = T_1$, in which case, $d_{NNI}(T_0, T_1) = 0, d_{NNI}(T_0, T_2) = 1$.

Case 2: $M_1 = A$. So, we have that T_1 contains the split $M_1B|CD$ and T_2 contains the split $BC \mid M_1D$. We have three possibilities for T_0, namely, it could contain one of the following three splits: $M_1B \mid CD$, $BC \mid M_1D$, or $BD \mid M_1C$. In each of these cases, we have $d_{NNI}(T_0, T_1), d_{NNI}(T_0, T_2) \leq 1$ and the lemma holds.

Case 3: $M_1 \supsetneq A$. So, $M_1 \cap B \neq \emptyset$. Since T_{M_1} is a subtree of T_1 and of T_2, it must contain all of B. If $M_1 = A \cup B$, then the target edges in T_1 and T_2 must separate C and D, and are identical. Similarly, if $M_1 \subsetneq A \cup B$, M_1 must contain all of C or all of D, and the taget edges in T_1 and T_2 must preserve the rooting of the remaining subtree, and thus, are identical. Thus, $d_{NNI}(T_0, T_1), d_{NNI}(T_0, T_2) = 0$.

□

Lemma 14. *Let $U \subseteq O_e$ be connected consist of trees more than 2 NNI moves from T_0. Let $n = |U|$. Then any NNI-circuit of U takes at least $\frac{3}{2}(n-1)$ steps.*

Proof. By induction on the size of $|U|$.

For $|U| = 1$: This is trivially true.

For $|U| > 1$, choose $x \in U$ closest to T_0. Since x is closest to T_0, not all of the neighbors of x are in U (if so, then there is an element in U closer to T_0). Since T_0 is binary, it has at most 4 neighbors in N_{SPR}. If x has one neighbor in U, then, a circuit of U must traverse the same edge from x to its neighbor twice, and the number of steps needed is at least two more than the number of steps needed for the smaller set $|U| - \{x\}$. By inductive hypothesis, this smaller set takes at least $\frac{3}{2}(|U - \{x\}| - 1)$ steps. So, the number of steps for U is:

$$\frac{3}{2}(|U - \{x\}| - 1) + 2 \geq \frac{3}{2}(|U| - 1)$$

If x has two neighbors in U, then call the subtrees rooted at x's neighbors, U_1 and U_2. If the neighbors of x are connected, then it takes 3 steps to visit x in a circuit of x, U_1, and U_2. If they are not connected, it takes 4 steps. Thus, by inductive hypothesis, the number of steps needed is:

$$\frac{3}{2}(|U_1| - 1) + \frac{3}{2}(|U_2| - 1) + 3 \geq \frac{3}{2}(|U| - 1)$$

If x has 3 neighbors in U, then by similar argument, we have the lower bound.

If x has 4 neighbors in U, then it is not the closest element of U to T_0, giving a contradiction. □

From the last two lemmas, we have that the orbits are mostly isolated– the only trees having neighbors from outside the orbit are within 2 steps of T_0. Each of these isolated arms of the orbit must be visited in an NNI walk of the SPR neighborhood, and the walks of the isolated arms take many extra steps. This yields our lower bound:

Lemma 15. *It takes $\Omega(n)$ extra steps to make a circuit of an orbit.*

Proof. Let $e \in E(T_0)$ and O_e its orbit. Since each orbit has $2n - 5$ trees (Observation 10) and by Lemma 11, at most 8 have neighbors from $N_{SPR} - O_e$.

It follows from Lemma 11, the $2n - 13$ remaining trees are in two connected sets. By the Pigeonhole Principal, one set has at least $n - 7$ trees. By Lemma 14, it takes $\Omega(((n-7)-1)/2) = \Omega(n)$ extra steps to visit the larger connected set.

Thus, it takes $\Omega(n)$ extra steps to traverse the orbit. □

The above lemmas can be combined to show that $\theta(n^2)$ extra steps are needed to traverse the neighborhood, since there are $2n - 3$ orbits, and each has minimal overlap with other orbits.

Theorem 16. *Every SPR neighborhood takes $(2n - 6)(2n - 7) + \Theta(n^2)$ steps to traverse.*

Proof. The upper bound follows by Lemma 12.

For the lower bound: by Lemma 13, every orbit, O_e has $\Omega(n)$ trees that have no neighbors in other orbits. By Lemma 15, it takes $\Omega(n)$ extra steps to traverse these regions of O_e. Since, by Theorem 9, there are $2n - 3$ orbits, we have that any path must take $\geq (2n-3)\Omega(n) = \Omega(n^2)$ extra steps. □

4 Acknowledgments

For spirited discussions, we would like to thank the members of the Treespace Working Group at CUNY: Ann Marie Alcocer, Kadian Brown, Samantha Daley, John DeJesus, Eric Ford, Michael Hintze, Daniele Ippolito, Joan Marc, Oliver Mendez, Diquan Moore, and Rachel Spratt.

This work was supported by grants from the US National Science Foundation programs in mathematical biology and computational mathematics (NSF #09-20920) and the New York City Louis Stokes Alliance for Minority Participation in Research (NSF #07-03449). Some of this work was completed at the Isaac Newton Institute for Mathematical Sciences at Cambridege University, and we gratefully acknowledge travel funding to the last author provided by the institute and the Lehman College-IBM Research travel fund.

References

[1] B. Allen and M. Steel. Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of Combinatorics*, 5:1–13, 2001.

[2] Maria Luisa Bonet, Katherine St. John, Ruchi Mahindru, and Nina Amenta. Approximating subtree distances between phylogenies. *Journal of Computational Biology*, 13(8):1419–1434 (electronic), 2006.

[3] Magnus Bordewich, Catherine McCartin, and Charles Semple. A 3-approximation algorithm for the subtree distance between phylogenies. *Journal of Discrete Algorithms*, 6(3):458–471, 2008.

[4] Magnus Bordewich and Charles Semple. On the computational complexity of the rooted subtree prune and regraft distance. *Annals of Combintorics*, 8:409–423, 2004.

[5] David Bryant. Kaikoura 2009 Penny Ante: A mathematical challenge. http://www.math.canterbury.ac.nz/bio/events/kaikoura09/penny.shtml, 2008.

[6] Alan Joseph J. Caceres, Samantha Daley, John DeJesus, Michael Hintze, Diquan Moore, and Katherine St. John. Walks in phylogenetic treespace. *Information Processing Letter*, 111:600–604, 2011.

[7] Bhaskar DasGupta, Xin He, Tao Jiang, Ming Li, John Tromp, and Louxin Zhang. On computing the nearest neighbor interchange distance. In D.Z. Du, P.M. Pardalos, and J. Wang, editors, *Proceedings of the DIMACS Workshop on Discrete Problems with Medical Applications*, volume 55 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pages 125–143. American Mathematical Society, 2000.

[8] L. R. Foulds and R. L. Graham. The Steiner problem in phylogeny is NP-complete. *Adv. in Appl. Math.*, 3(1):43–49, 1982.

[9] Glenn Hickey, Frank Dehne, Andrew Rau-Chaplin, and Christian Blouin. SPR distance computation for unrooted trees. *Evolutionary Bioinformatics*, 4:17–27, 2008.

[10] D. M. Hillis, B. K. Mable, and C. Moritz. *Molecular Systematics.* Sinauer Assoc., Sunderland, Mass., 1996.

[11] Ming Li, John Tromp, and Louxin Zhang. Some notes on the nearest neighbour interchange distance. In *COCOON '96: Proceedings of the Second Annual International Conference on Computing and Combinatorics*, pages 343–351, London, UK, 1996. Springer-Verlag.

[12] Sebastian Roch. A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(1):92–94, 2006.

[13] Charles Semple and Mike Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications.* Oxford University Press, Oxford, 2003.

[14] Mike Steel. Challenges and Conjectures: Isaac Newton Institute June 2011. http://www.newton.ac.uk/programmes/PLG/phylogenetics_challenges.pdf.

[15] Simon Whelan. New approaches to phylogenetic tree search and their application to large numbers of protein alignments. *Syst. Biol.*, 56(5):727–740, 2007.